DOCUMENT RESUME

ED 285 902 TM 870 488

AUTHOR Sandler, Andrew B.
TITLE The Use of Invariance and Bootstrap Procedures as a Method to Establish the Reliability of Research Results.
PUB DATE 30 Jan 87
NOTE 19p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 29-31, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Covariance; Analysis of Variance; Correlation; Discriminant Analysis; *Generalizability Theory; *Mathematical Models; Regression (Statistics); *Reliability; Research Design; Research Problems; *Sample Size; Sampling; Simulation; Statistical Inference; *Statistical Significance; Statistical Studies; Validity
IDENTIFIERS *Bootstrap Hypothesis; *Cross Validation; Invariance Principle

ABSTRACT

Statistical significance is misused in educational and psychological research when it is applied as a method to establish the reliability of research results. Other techniques have been developed which can be correctly utilized to establish the generalizability of findings. Methods that do provide such estimates are known as invariance or cross-validation procedures and the bootstrap method. Invariance procedures split the total sample into two subgroups and apply techniques to analyze each subgroup and compare results, often by using parameters obtained from one subgroup to evaluate the other subgroup. A simulated data set is presented and analyzed by invariance procedures for: (1) canonical correlation; (2) regression and discriminant analysis; (3) analysis of variance and covariance; and (4) bivariate correlation. Whereas invariance procedures split a sample into two parts, the bootstrap method creates multiple copies of the data set. The number of copies could exceed millions with current computer capability. The copies are shuffled and artificial samples of 20 cases each, called bootstrap samples, are randomly selected. The value of the Pearson product-moment correlation (or other statistics) is then calculated for each bootstrap sample to assess the generalizability of the results. (LPG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED285902

THE USE OF INVARIANCE AND BOOTSTRAP PROCEDURES AS A METHOD TO ESTABLISH THE RELIABLITY OF RESEARCH RESULTS:

Andrew B. Sandler

University of New Orleans

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, Jan. 30,1987

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A.B. Sandler

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 870 488

BEST COPY AVAILABLE

ABSTRACT

Statistical significance testing is misused in research in education and psychology when it is applied as a method to establish the reliability of research results. Other techniques have been developed which can be correctly utilized to establish the generalizibility of research results. Methods that do provide such estimates are known as invariance or cross-validation procedures and the bootstrap method. These two procedures are discussed in this paper and concrete examples are presented.

Statistical significance testing is a common component of research in education and psychology. The emphasis on statistical tests is reflected by the fact journal editors may reject articles if these techniques are not included. Although there are alternative methods which can be used to analyze and summarize data; the use of traditional hypothesis-testing inferential statistics is the only tool that most researchers know how to utilize when they conduct their studies.

Several researchers have questioned the use of statistical significance testing as the primary means of interpreting data. Carver (1978, p.378) argues:

> Statistical significance testing has involved more fantasy than fact. The emphasis on statistical significance in educational research represents a corrupt form of the scientic method. Educational research would be better off if it stopped testing its results for statistical significance.

One problem with statistical significance testing is that statistical significance ordinarily depends upon how many subjects are used in the research. Carver (1978, p.388) notes:

> Because statistical significance in research ordinarily depends on sample size, trivial results are often interpreted as "significant" or important when they are simply results that would rarely happen when randomly sampling from the same population using large

sample sizes. A mean difference that is small and not significant from a research standpoint can be statistically significant just because enough subjects were used in the experiment to make the result statistically rare under the nulll hypothesis.

Thompson (1986, pp.12-13) argues that:

> Many researchers try to employ large samples in order to "wash out the effects of individual differences or outliers, or to invoke the central limit theorem to meet distribution assumptions of parametric tests, or to maximize generalization... The significance test [often] only confirms what the researcher already knew, i.e. that he or she began the study with a large sample size.

The mathematical dependence of statistical significance upon sample size can make negligible research results appear "important". Lykken (1968) warned that "[finding statistical significance] is never a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence--or that an experimental report ought to be published" (p.158). Conversly, the failure to detect "significant" effects frequently is interpreted by many scientists to mean that the research is unimportant, even when the population size is itself very small (e.g; individuals suffering from rare diseases.)

Valuable research uses methods which honor the reality about which the researcher is attempting to generalize. In contemporary practice poorly designed studies (e.g; those which are vulnerable to extraneous variables, those with poor sampling techniques) are sometimes cited in an effort to support a theory because the results are significant even though they can not be generalized from a sample to a larger population. Ideally, researchers want to generalize from results that are consistent over many samples. The reliability or replicability of an experiment is one way to estimate consistency of research findings.

Carver (1978) has argued that statistical significance is not an index of reproducibility. The confounding of statistical significance with reproducibility is a major prevailing misconception about significance testing: Statistical significance at the .05 level does not necessarily imply that an investigator can be confident with odds of 95 out of 100 that the observed difference will hold up in future investigations.

Researchers who want to eliminate their dependence on statistical significance and still consider chance and reliability can search for replicated evidence by repeating the study with as many samples as possible. Unfortunately, this is rarely feasible. However, statistical procedures have been developed which generate data sets from the original data so that the actual variability of the sample results can be assessed. Two of these procedures are

invariance or cross-validation procedures and the "bootstrap" method. The purpose of this paper is to briefly discuss these methods. Table 1 presents a hypothetical data set (n=20) used to make the presentation of these two procedures concrete.

INSERT TABLE 1 ABOUT HERE.

The X and Y variables in Table 1 represent hypothetical intervally-scaled dependent variables. The A and B variables represent hypothetical intervally-scaled independent variables. The A1 and B1 variables are derived by converting A and B to the nominal level of scale in the same fashion that many OVA researchers treat aptitude variables in aptitude-treatment interaction OVA designs. The variable labeled INV divides the 20 cases into two groups. Group 0 contains 8 cases and group 1 contains 12 cases.

Invariance Procedures

Invariance procedures attempt to assess the stability of statistical results across different samples. In this procedure, an analysis is performed separately on each of two subgroups into which the sample has been divided. The results of each subgroup are analyzed and compared with one another. When the results of an analysis are comparable, the researcher can make more confident assumptions about the experimental hypothesis. When the results of an analysis

are not comparable (i.e., not invariant) then the researcher may have serious doubts about the generalizability of the results.

This paper will demonstrate that invariance procedures can be employed with a variety of parametric procedures. Although the details of invariance analysis vary according to the analytic technique employed, the logic of invariance remains consistent.

Canonical Analysis

Canonical correlation is the appropriate analytic technique when each variable set--predictor (independent) and criterion(dependent) variables--has two or more elements. Invariance procedures in canonical correlation analysis have been proposed by Thorndike (1978) and Thompson (1982). In the present example, criterion variables X and Y and predictor variables A and B were used to demonstrate one canonical invariance procedure. The invariance procedure for canonical analysis consists of the following steps:

1) The original sample is randomly divided into two invariance groups.

2) Within each group, all variables are separately standardized into Z-scores and separate canonical correlation analyses are computed for each of the two invariance groups.

3) The invariance of the canonical function is computed for group O. This is accomplished by computing a second set

of composite scores by using the data from group 0 and the standardized function coefficients from group 1.

4) The invariance of the canonical function is computed for group 1. This is accomplished by computing a second set of composite scores by using the data from group 1 and the standardized function coefficients from group 0.

Table 2 presents the actual and invariance canonical correlation coefficients used in this analysis.

Thompson (1981) defines invariance coefficients as estimates of degree of capitalization on sampling error. In this example, the actual canonical correlation coefficients for group 0 and group 1 were respectively, .7087 and .2779. When the weights from the analysis for group 1 were applied to data for group 0, the actual R^2c for group 0 (.7087) "shrunk" to .6960. The corresponding shrinkage for group 1 was from .2779 to .2736. The difference between the squares of the "genuine" and the cross-validated coefficient is an invariance estimate. The closer this value is to zero, the more stable the regression results may be assumed to be across samples. In this example, the invariance estimates--.012 for group 0 and .004 for group 1 are impressive (especially when the extremely small sample size is considered). The invariance estimate can be converted into a 0.0 to 1.0 metric by dividing the shrinkage by the "genuine" R^2c for an analysis, e.g., .012/.7087.

INSERT TABLE 2 ABOUT HERE

In this hypothetical study, the researcher would conclude that the results are replicable. Similar analyses could be conducted for function II. If the invariance estimate which was yielded for function I was not as close to zero then the researcher would have to consider the possibility that the results of this study do not generalize to the larger population. Of course, it is troublesome that the two "genuine" squared canonical coefficients were so different. It should be noted, that invariance analysis does not have anything to do with the interpretation of the study results. Even with high invariance estimates, large effect sizes still indicate that the variables included in an analysis did have a significant effect on the behavior of the study sample.

Regression and Discriminant Analysis

Researchers have developed methods for estimating the invariance of regression results (Huck, Cormier, and Bounds, 1974, p.159) and the invariance of discriminant analysis results (Huck, Cormier, and Bounds, 1974, p.174). The invariance procedures for these two methods are analogous in almost every detail to the procedure discussed above for canonical correlation analysis.

Table 3 provides the invariance estimates which were used in a regression invariance example. In this example, variable X from table 1 was used as the criterion variable and variables A and B were used as the predictor variables.

In this case an invariance estimate is the 0.0 to 1.0 metric and can also be computed by correlating for example, each group 0 subject's "genuine" predicted X score with each group 0 subject's predicted score produced using group 1's beta weights.

INSERT TABLE 3 ABOUT HERE

Table 4 provides the invariance estimates which were used in a discriminant analysis invariance example. In this example, variable A1 was used as a dichotomous dependent variable. Variable X and Variable Y were used as the predictor variables.

INSERT TABLE 4 ABOUT HERE

OVA Techniques

Thompson (1986) notes that ANOVA and ANCOVA are special cases of multiple regression and that all parametric methods are special cases of canonical correlation analysis. Thompson suggests that invariance estimation procedures might also be useful with OVA analyses.

Invariance procedures with OVA methods are thusly done:

1) The original sample is randomly divided into two invariance groups.

2) The effect size (the sum of squares for an effect is divided by the sum of squares total) is estimated.

3) The effect size for each effect is compared between invariance groups.

Although many researchers who are concerned with significance testing may prefer to compare significance levels instead of effect sizes, Thompson (1986) argues that "Effect sizes represent an appropriate context for result interpretation." Other more sophisticated effect sizes (e.g. omega squared) can also be compared.

Table 5 provides Wilks's Lambda for invariance group 0 and invariance group 1 in a MANOVA invariance example. In this example Variable X and Variable Y were used as the criterion variables and A1 and A2 were used as the ways. The reader should recall that Wilks's Lambda is a ratio which indicates the proportion of total sum of squares which is not accounted for by the treatment effect. By subtracting Wilks's Lambda from 1.0, we obtain the statistic analogous to effect size.

INSERT TABLE 5 ABOUT HERE

Table 6 provides the effect sizes which were computed from invariance group 0 and invariance group 1 in a factorial ANOVA. In this example, variable X was considered the criterion variable and A1 and B1 as the ways in the design. This is an example of a 2x2 factorial ANOVA.

INSERT TABLE 6 ABOUT HERE

Table 7 provides the effect sizes which were computed from invariance group 0 and invariance group 1 in a 1 way ANOVA. In this example, variable X was considered the criterion variable and A1 as the way in the design.

INSERT TABLE 7 ABOUT HERE

Bivariate Correlation

Table 8 provides the squared correlation coefficients which were computed from invariance group 0 and invariance group 1 in a Pearson product moment correlation. In this example, variable X was correlated with variable Y.

INSERT TABLE 8 ABOUT HERE

Bootstrap Procedures:

The bootstrap procedure (Diaconis & Efron, 1983) is another way in which the statistical accuracy of the data in a single sample can be assessed. This is accomplished by generating many samples from the data in the original sample. The bootstrap method could be used to evaluate a correlation between variable X and Y in all 20 cases in our hypothetical data set:

1) The data from each case are copied a large number of times (with computers each case could be copied a million times). All 20 million copies are thoroughly shuffled.

2) Artifical samples of the 20 cases, called bootstrap samples, are created by selecting sets of 20 data points at random from the 20 million cases.

3) The value of r is then calculated for each of the bootstrap samples.

The bootstrap procedure can also be used with other statistical techniques besides the Pearson product moment correlation. This technique has the advantage of generating larger sample sizes than the invariance procedures which are dependent on the split of the original sample. Diaconis and Efron (1983) note that the bootstrap method is a new phenomenon which was not made possible until the development of high speed computers. Thus,these authors note the bootstrap technique needs theoretical justification on complex statistical problems.

Summary

Researchers always need to address the question of whether their obtained results will generalize to the larger population. Although popular, statistical significance tests fail to provide estimates for the stability of research results. Invariance techniques and the bootstrap method are useful ways in which the generalizability of research can be determined.

References

Carver R.C. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Diaconis, P. & Efron,B. (1983). Computer-intensive methods in statistics. Scientific American, 24, 116-130.

Huck, S.W., Cormier, W.H. & Bounds, W.G., jr. Reading statistics in research. New York: Harper & Row, 1974.

Lykken, D.T. (1968). Statistical significance in psychological research. Psychological Bulletin, 70, 151-159.

Thompson, B. (1981). Utility of invariance coefficients. Perceptual and Motor skills, 52, 708-710.

Thompson, B. (1986). Anova versus regression analysis of ATI designs: An empirical investigation. Educational and Psychological Measurement, 46, 917-928.

Thompson, B. (April, 1986). The place of qualitative research in contemporary social science: The importance of post-paradigmatic thought. Paper presented at the anual meeting of the American Research Association, San Francisco.

Table 1
Hypothetical data set

Case:	X	Y	A	B	A1	BI	INV
1	6	3	2	3	1	1	0
2	1	2	3	4	1	1	0
3	8	3	3	7	1	2	0
4	9	4	4	8	1	2	0
5	1	8	9	4	2	1	0
6	3	6	8	3	2	1	0
7	2	1	8	7	2	2	0
8	8	5	9	6	2	2	0
9	2	0	3	2	1	1	1
10	7	7	4	3	1	1	1
11	5	8	2	1	1	1	1
12	6	8	3	7	1	2	1
13	4	6	2	6	1	2	1
14	3	7	1	8	1	2	1
15	9	3	8	3	2	1	1
16	5	6	7	2	2	1	1
17	6	6	6	4	2	1	1
18	5	2	7	6	2	2	1
19	2	8	9	8	2	2	1
20	7	4	8	7	2	2	1

table 2

Squares of actual and Invariance Canonical Correlation Coefficients for function 1

Group 0 data and group 0 Canonical coefficients = .7087
Group 0 data and group 1 Canonical coefficients = .6960
Group 1 data and group 1 Canonical coefficients = .2779
Group 1 data and group 0 Canonical coefficients = .2736

Invariance estimate for group 0 : .7087-.6906 = .012
Invariance estimate for group 1 : .2779-.2737 = .004

table 3

Squares of actual and Invariance Canonical Correlation Coefficients

Group 0 data and group 0 regression coefficients = .3930
Group 0 data and group 1 regression coefficients = .3410
Group 1 data and group 1 regression coefficients = .1396
Group 1 data and group 0 regression coefficients = .1084

Invariance estimate for group 0: .3930-.3410 = .052
Invariance estimate for group 1: .1396-.1084 = .0312

table 4

Squares of actual and invariance discriminant correlation coefficients for function 1

Group 0 data and group 0 discriminant coefficient = .3533
Group 0 data and group 1 discriminant coefficient = .3410
Group 1 data and group 1 discriminant coefficient = .1349
Group 1 data and group 0 discriminnat coefficient = .0545

Invariance estimate for group 0: .3533-.3410 = 0123
Invariance estimate for group 1: .1349-.0545 =.0804

table 5

Manova Lambda values

Invariance group 0

	Lambda
A1	.77
B1	.807
A1xB1	.6108

Invariance group 1

	Lambda
A1	.1510
B1	.1262
A1xB1	.1000

table 6
Factorial Anova

Invariance group 0

	SS	Effect Size
A1	12.5	.1572
B1	32	.4025
A1xB1	2	.0251
total	79.5	

Invariance group 1

	SS	Effect Size
A1	4.08	.0834
B1	4.083	.0934
A1xB1	2.083	.0425
total	48.917	

table 7
1-way Anova

Invariance group 0

	SS	Effect Size
A	12.5	.1572
total	79.5	

Invariance group 1

	SS	Effect Size
A	4.083	.0834
total	48.917	

table 8
Squared Pearson Correlation

Invariance group 0 = .0055
Invariance group 1 = .0005